Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W615–W619***
DOI: 10.1093/nar/gkh398

MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes

Ian W. Davis, Laura Weston Murray, Jane S. Richardson and David C. Richardson*

Department of Biochemistry, Duke University, Durham, NC 27710-3711, USA

Received February 20, 2004; Revised and Accepted March 25, 2004

ABSTRACT

MolProbity is a general-purpose web service offering quality validation for three-dimensional (3D) structures of proteins, nucleic acids and complexes. It provides detailed all-atom contact analysis of any steric problems within the molecules and can calculate and display the H-bond and van der Waals contacts in the interfaces between components. An integral step in the process is the addition and full optimization of all hydrogen atoms, both polar and nonpolar. The results are reported in multiple forms: as overall numeric scores, as lists, as downloadable PDB and graphics files, and most notably as informative, manipulable 3D kinemage graphics shown on-line in the KiNG viewer. This service is available free to all users at http://kinemage.biochem.duke.edu.

INTRODUCTION

Experimental structure determination has rather different strengths and weaknesses for nucleic acids than for the protein case, where most validation methods were developed. Positions and interactions of the bases can be quite accurately determined, but for both X-ray and NMR methods, much of the sugar–phosphate backbone is quite difficult and ambiguous, with too many degrees of freedom relative to the observable data. Figure 1 contrasts the reproducibly well-fit all-atom contacts of RNA bases with the frequent steric clashes of H-atoms seen in RNA backbone in the 2.5–3 Å resolution range typically attained for large, biologically important nucleic acids. Structural biologists fully appreciate the difficulty with backbone, but so far have lacked good tools for diagnosis or remediation. While existing torsion angle analyses (1,2) are substantially correct, errors in one or more angles, resulting in impossible conformations, are common (3,4). Traditional clash analysis tools (5,6) do not use the hydrogens, which are especially revealing in this case. The all-atom contact analysis (7) featured on the MolProbity site provides a simple but powerful diagnostic tool for nucleic acid backbone, and its local and directional nature can even suggest how to make improvements. That same analysis gives the end-users of nucleic acid structures an easy way to assess local accuracy in a region of interest. An all-atom contact visualization of the interface between two molecules also gives a direct, intuitive way to see the H-bond and van der Waals interactions.

METHODS

MolProbity is implemented using the scripting language PHP in conjunction with the Apache web server. External programs written in C, Java and other languages are invoked by MolProbity to analyze the structures and generate kinemage visualizations. The MolProbity PHP code collects and parses the output of these programs and presents the results in a meaningful way. PHP code is also responsible for creating the user interface (in the form of web pages), controlling program flow (e.g. which tools are available when) and managing user data over the lifetime of a session.

Input is a PDB-format macromolecular coordinate file from the Protein Data Bank (8) or the Nucleic Acid Database (9), or can be uploaded from the results of a structure determination. We identify structures used in the examples here by both PDB and NDB codes (e.g. 1JJ2/rr0033). All hydrogen atoms, both polar and nonpolar, are added by the Reduce program (10). Reduce's expert system uses the information from both hydrogen bonding and all-atom steric compatibility to fully optimize local H-bond networks and correct 180° 'flips' for Asn, Gln and His orientations in the proteins. Base tautomers are not varied, and only the first layer of waters is considered, to minimize sensitivity to errors in positioning and to keep the H-bond networks small enough for deterministic analysis. This step produces a commented, modified PDB file and also a graphic display of the consequences of each proposed side-chain flip; any changes deemed unacceptable can be overridden by the user.

With all hydrogens present, all-atom contacts are then calculated by Probe (7), which uses traditional van der Waals radii (11) for most atoms and 1.0 Å for polar H, in a rolling-probe algorithm that leaves a dot when the 0.25 Å-radius probe

*To whom correspondence should be addressed: Tel: +1 919 684 6010; Fax: +1 919 684 8885; Email: dcr@kinemage.biochem.duke.edu

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

© 2004, the authors

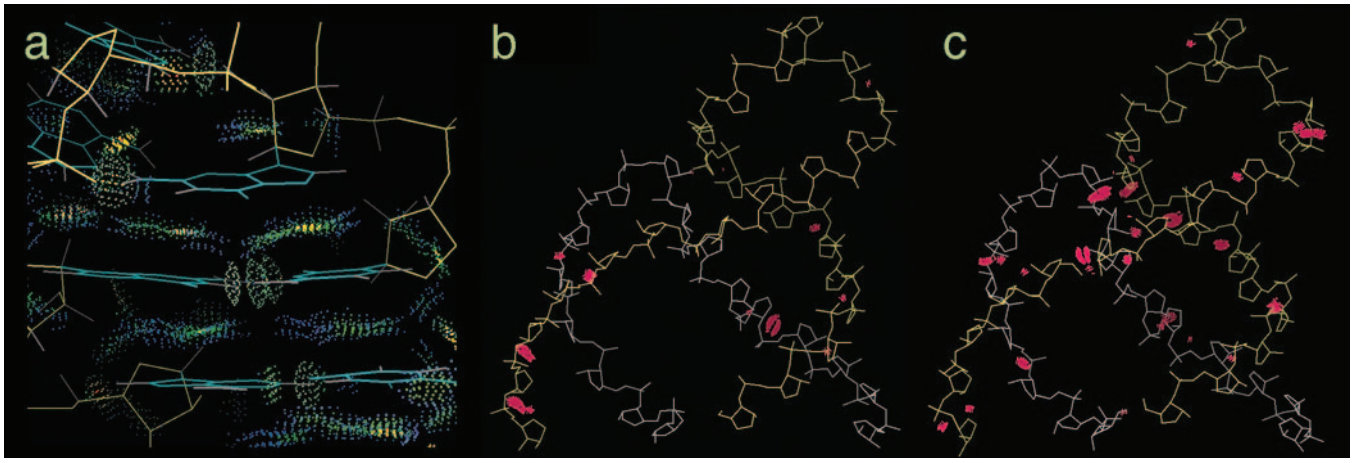

Figure 1. All-atom contacts used to assess structure quality for hammerhead ribozymes, as calculated in MolProbity and displayed online in the KiNG Java viewer. (**a**) Well-fit base–base and base–backbone interactions (see Methods for color code) in the tetraloop region of the 359D/urx067 structure (14). (**b**) Overview of the 359D/urx067 backbone, showing just the serious steric clashes (red spikes); its clash score is 55. (**c**) Overview of the similar 379D/urx071 backbone (19), showing just the serious clashes; its clash score is 139, and the clashes are not in the same positions.

intersects another not-covalently-bonded atom (7). The results include a clustered list of disallowed atom pair overlaps ⩾0.4 Å, an overall clash score (number of bad overlaps per 1000 atoms) and two kinemage graphics displays (12,13) of contacts for the whole structure. Van der Waals contacts are shown as back-to-back patches of green or blue dots on the surfaces of non-covalent atom pairs within 0.5 Å of touching (as in Figure 1a); hydrogen bonds are shown as lenses of pale green dots outlining the interpenetrating surfaces of a donor and acceptor. Steric clash overlaps are emphasized with spikes rather than dots, progressing from yellow to hot pink as the clash becomes truly serious beyond 0.4 Å overlap (Figure 1b).

The graphics in MolProbity are displayed interactively online in the KiNG Java kinemage viewer. KiNG can smoothly rotate very large structures, with extensive user control of display and measurement options, and it shows all of the contact dot surfaces, H-atoms, animations, ribbons, two-dimensional (2D) plots and so on that are produced by the MolProbity analyses. Electron density maps (e.g. from the Electron Density Server at http://fsrv1.bmc.uu.se/eds/) can be uploaded and added to the display. Alternatively, the various kinemage files can be downloaded and shown off-line either in KiNG or in Mage (12,13). Those, and standalone versions of the other programs used in MolProbity, are available as free, open-source, multi-platform software from http://kinemage.biochem.duke.edu; it is also possible to install a copy of MolProbity on a local computer.

MODE OF USE

MolProbity is accessed from the main navigation bar on the kinemage website at http://kinemage.biochem.duke.edu. The interactive graphics with KiNG require Java version 1.3 or higher. The user specifies a PDB identification code or uploads a coordinate file for input. A short summary of the structure (name, resolution, number of chains and residues, etc.) is presented, along with a rotatable thumbnail image, to verify the file's identity and contents. Both a MolProbity tutorial and a user manual are provided. The main page, as shown in Figure 2, then offers a menu of options, some of which are specific to proteins (e.g. Ramachandran plot and side chain rotamer evaluations). Unless both nonpolar and polar hydrogens are already present in the PDB file, nucleic acid analysis begins with H-atom addition, which usually takes just a few seconds but can take 15 min for a ribosome structure. The file with all H-atoms added and optimized can be downloaded at any time during the session. With all atoms present, the various all-atom contact analyses are now available.

The all-atom contact function brings up a dialog box that controls a Probe run on the structure. The default setting calculates steric overlaps to locate problem areas, which is quite fast and feasible even on extremely large structures. For most structures, it is advisable to check the boxes for calculating H-bond and van der Waals contacts as well. Two kinemage files are produced, one with the backbone–backbone contacts and one with base–base and base–backbone contacts, each along with a complete display of the molecule(s). The side chain contacts are the most important for protein evaluation, but the backbone contacts are the dominant validation criteria for nucleic acid structures (since the base contacts are usually good). This information can be studied online in the KiNG Java kinemage viewer. An area with extensive dot surfaces in the favorable green, blue and yellow colors, such as the tetraloop base-stack of the hammerhead ribozyme structure 359D/urx067 (14) in Figure 1a, is very well positioned and reliable. When only the bad contacts are displayed, as for the two structures compared in Figures 1b and 1c, then it is easy to find the clumps of red spikes that mean something must be wrong in that part of the model. Note that although the resolutions are similar (2.9 versus 3.1 Å), one structure has more than twice as many clashes as the other (clash scores 55 versus 139). Also, that the clashes are in different places underscores the fact that they represent local misfittings. A structural biologist can zoom in on each such position and look for a way to improve its local conformation. A molecular biologist, bioinformaticist or other person wanting to extract information from the molecule can easily judge structure quality for their region of interest: if the area is free of

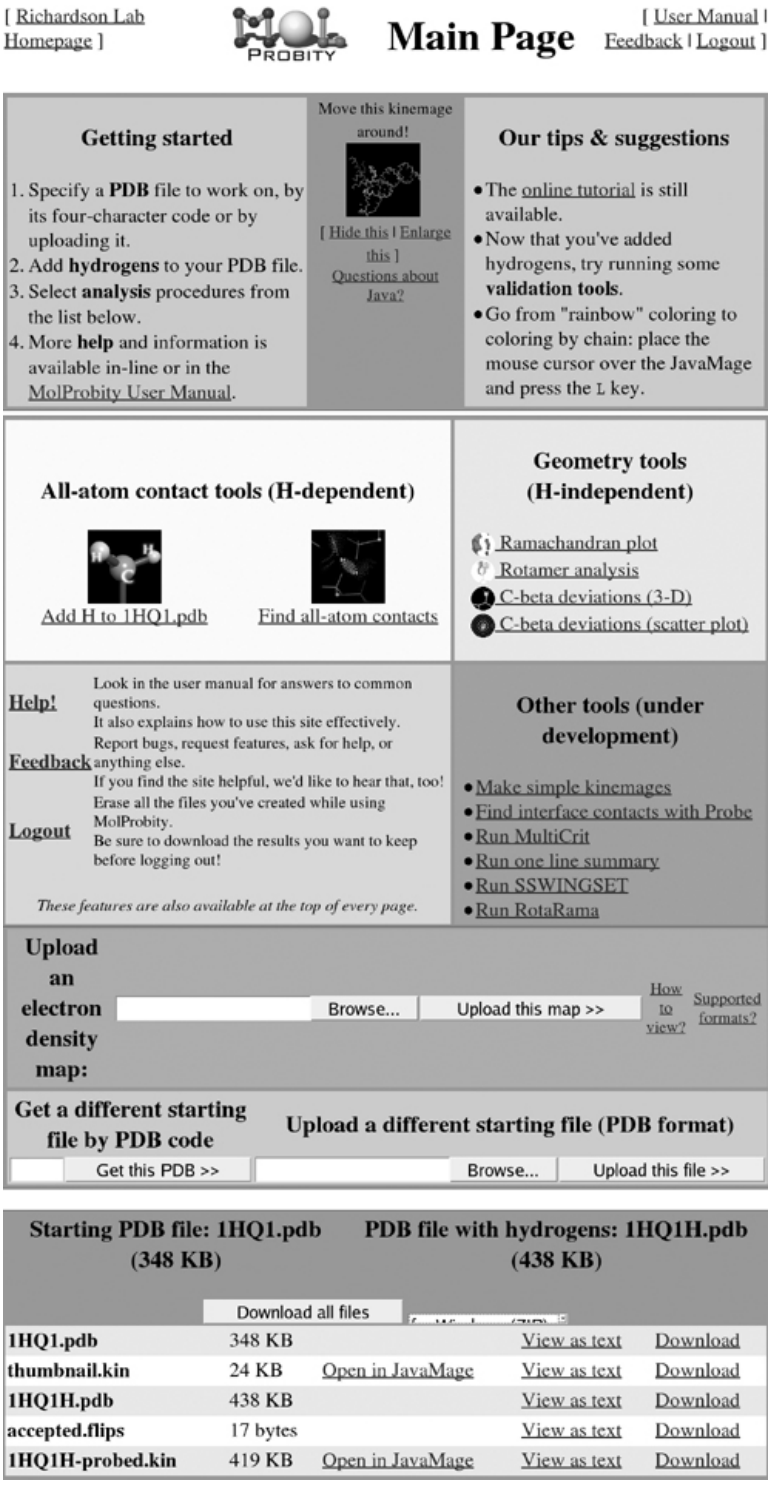

Figure 2. The main page of the MolProbity server during a session analyzing the 1HQ1/pr0037 protein–RNA complex (20), with a rotatable thumbnail image at the top, contact and geometry analysis tools in the center and control of file uploads and downloads at the bottom.

clashes, then it should be accurate even at a quite detailed level. For those who prefer lists to images, the clashlist option gives a list of all serious steric clashes, plus an overall clash score (serious clashes per 1000 atoms). Even in structures at atomic resolution, some clashes still occur, but they are of course very much rarer.

As well as diagnosing model problems, all-atom contacts are an excellent way to study molecular contacts. Under 'Other Tools' (Figure 2), MolProbity provides an interface for specifying the calculation and display of all atom–atom contacts in the interface between different chains or groups in the input file, such as between a drug and DNA, or between a protein

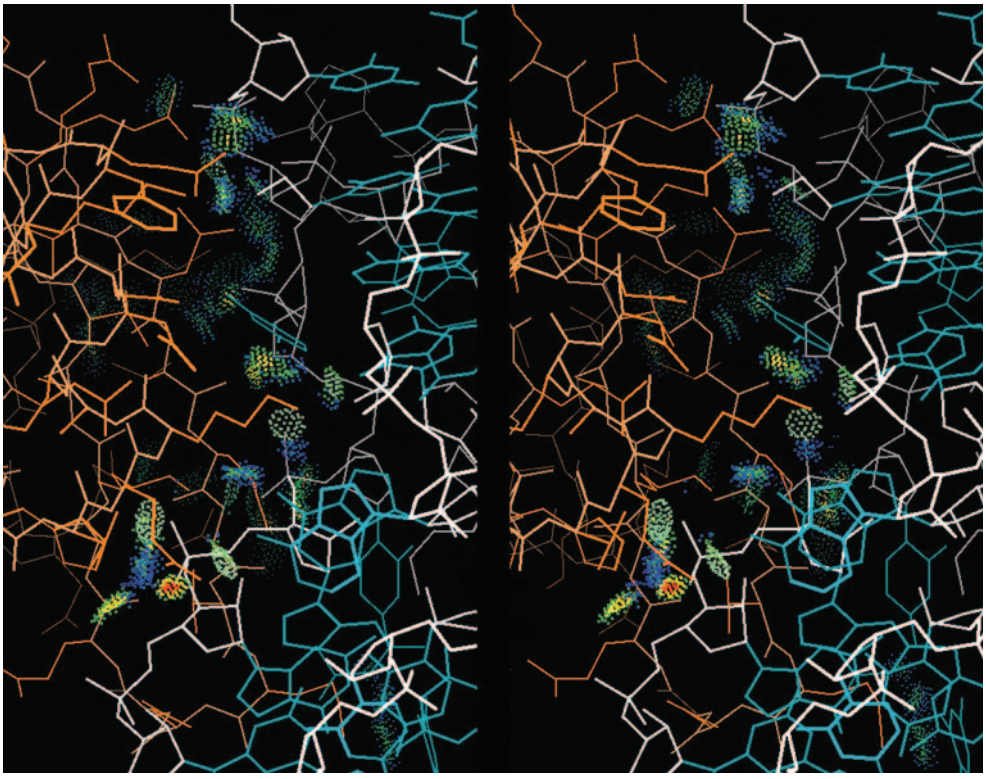

Figure 3. All-atom contacts for the interface between ribosomal protein L23 (on left, in peach) and a four-way junction region of the 23S RNA (on right, with white backbone and cyan bases), from the 1JJ2/rr0033 structure of the 50S ribosomal subunit from *Haloarcula xmarismortui* (15). Only two bases are involved, almost all contact being between protein side chains and the RNA backbone. (See Methods for color coding of the dot surfaces.)

and RNA, as illustrated in Figure 3 for the interactions of protein L23 with RNA in the 1JJ2/rr0033 structure of the 50S ribosomal subunit (15). The hydrogen bonds and the van der Waals contact displays can be turned on or off separately, and the specific atoms involved can be identified either visually or by clicking on a dot. Interestingly, only two bases (U1447 and A1501) are involved directly; this contact achieves high specificity by many interactions (e.g. 14 of the 15 H-bonds) recognizing the unusual RNA backbone conformation of this four-way junction.

An additional MolProbity option is to make simple, prescripted kinemages of the structure. The two most useful forms for nucleic acids are ribbons and a grouped version with base type color-coded (G green, C yellow, A pink, U/T blue). The 'L' key toggles backbone coloring between a blue-to-red 5′ to 3′ spectrum and one pastel color per chain. Under the RotaRama option, a 2D contact map shows the all-atom contacts internal to the structure as a color-coded diagonal plot.

If the structure includes protein, that part can be analyzed with a set of geometrical validation tools (box at right center in Figure 2) that complement the all-atom contacts. Ramachandran plot criteria (16) and side chain rotamer distributions (17) have been updated and improved, and the deviation of C-beta positions from ideal (16) summarizes bond angle distortions around the sensitive C-alpha position. The 'multi-criterion' display shows the outliers in each of the four independent measures, all together on the three-dimensional (3D) structure.

The bottom box of the main page (Figure 2) accumulates result files over the course of a MolProbity work session. The final step for a user is to download the files desired, before logging out.

DISCUSSION

Correct details of nucleic acid 3D structure are biologically important for understanding the specificity and control of protein–DNA, protein–RNA, RNA–RNA and aptamer binding interactions, and especially for mechanistic understanding of RNA catalysis. The critical base-pairing interactions and the phosphate positions can be very well determined crystallographically, which fortunately keeps any problems quite local; however, the many rotatable backbone angles make such local misfittings very hard to avoid, and they are quite prevalent even in very carefully determined structures. All-atom contacts (including all hydrogens) provide new information that is independent of the target functions used in refinement and that can sensitively and easily locate such problems (4,7). We have shown, for proteins, that correction of these model errors improves the R_{work} and R_{free} measures of agreement with data as well as improving the geometry and sterics (18). Such analysis can now be applied to nucleic acids as well, using the MolProbity web service.

The all-atom contact tools are useful in different ways for two different audiences. On the one hand, structural biologists can check out a structure in the process of determination by crystallography or NMR and quickly locate the local regions that most need attention. We are working on tools to make the

correction process easier, but traditional methods often work once it is clear which atoms need to move. On the other hand, anyone seeking to mine information from structure coordinates already deposited in the NDB or PDB can quickly assess on-line the reliability of conformational details in the region of special interest to them.

Intermolecular contacts are central to the biological interpretation of macromolecular data. Many useful systems are already available for assigning and showing such contacts, usually in the form of lists or simple diagrams with yes/no decisions based on rather arbitrary cutoffs. This works extremely well for base-pair interactions. For more complex interfaces, all-atom contact displays have the advantage of explicitly highlighting which atoms touch, within the context of the 3D relationships shown at a controllable level of detail. MolProbity provides such contact displays between any combination of chains or 'het' groups in the file under analysis. More complex selections of regions or contact types can be done with stand alone versions of the programs available from the same web site.

MolProbity will continue to develop further features for analysis of nucleic acid structures. Most central will be the addition of geometrical validation tools for RNA and DNA, since the quality-filtered data analysis of RNA backbone (4) has shown that criteria for backbone conformers, sugar puckers, and so on should be definable to fill the same role for nucleic acids that Ramachandran plots and side chain rotamers do for proteins.

ACKNOWLEDGEMENTS

Research and implementation for the MolProbity server and the software it utilizes are supported by NIH grants GM-15000 (D.C.R.) and GM-61302 (J.S.R.) and by Howard Hughes Medical Institute fellowships (I.W.D. and L.W.M.).

REFERENCES

1. Lakshminarayanan,A.V. and Sasisekharan,V. (1970) Stereochemistry of nucleic acids and polynucleotides II. Allowed conformations of the monomer unit for different ribose puckerings. *Biochim. Biophys. Acta*, **204**, 49–59.
2. Parkinson,G., Vojtechovsky,J., Clowney,L., Brunger,A. and Berman,H.M. (1996) New parameters for the refinement of nucleic acid containing structures. *Acta Cryst.*, **D52**, 57–64.
3. Murthy,V.L. and Rose,G.D. (2003) RNABase: an annotated database of RNA structures. *Nucleic Acids Res.*, **31**, 502–504.
4. Murray,L.J.W., Arendall,W.B.,III, Richardson,D.C. and Richardson,J.S. (2003) RNA backbone is rotameric. *Proc. Natl Acad. Sci. USA*, **100**, 13904–13909.
5. Hooft,R.W.W., Sander,C. and Vriend,G. (1996) Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins*, **26**, 363–376.
6. Westbrook,J., Feng,Z.K., Burkhardt,K. and Berman,H.M. (2003) Validation of protein structures for protein data bank. *Methods Enzymol.*, **374**, 370–385.
7. Word,J.M., Lovell,S.C., LaBean,T.H., Taylor,H.C., Zalis,M.E., Presley,B.K., Richardson,J.S. and Richardson,D.C. (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogens. *J. Mol. Biol.*, **285**, 1711–1733.
8. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
9. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.-H., Srinivasan,A.R. and Schneider,B. (1992) The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
10. Word,J.M., Lovell,S.C., Richardson,J.S. and Richardson,D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**, 1735–1747.
11. Bondi,A. (1964) Van der Waals Volumes and Radii. *J. Phys. Chem.*, **68**, 441–451.
12. Richardson,D.C. and Richardson,J.S. (1992) The Kinemage: a tool for scientific illustration. *Protein Sci.*, **1**, 3–9.
13. Richardson,J.S. and Richardson,D.C. (2002) MAGE, PROBE, and Kinemages, chapter 25.2.8. In Rossmann,M.G. and Arnold,E. (eds), International Tables for Crystallography. Kluwer Academic Publishers, Dordrecht, The Netherlands. Vol. F, pp. 727–730.
14. Feig,A.L., Scott,W.G. and Uhlenbeck,O.C. (1998) Inhibition of the hammerhead ribozyme cleavage reaction by site-specific binding of Tb (III). *Science*, **279**, 81–84.
15. Klein,D.J., Schmeing,T.M., Moore,P.B. and Steitz,T.A. (2001) The kink-turn: a new RNA secondary structure motif. *EMBO J*, **20**, 4214–4221.
16. Lovell,S.C., Davis,I.W., Arendall,W.B., III, de Bakker,P.I.W., Word,J.M., Prisant,M.G., Richardson,J.S. and Richardson,D.C. (2003) Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins*, **50**, 437–450.
17. Lovell,S.C., Word,J.M., Richardson,J.S. and Richardson,D.C. (2000) The penultimate rotamer library. *Proteins*, **40**, 389–408.
18. Richardson,J.S., Arendall,W.B.,III., and Richardson,D.C. (2003) New tools and data for improving structures, using all-atom contacts. In Carter,C.W.Jr. and Sweet,R.M. (eds) Methods in Enzymology: Macromolecular Crystallography. Academic Press, New York, 374, pp. 385–412.
19. Murray,J.B., Terwey,D.P., Maloney,L., Karpeisky,A., Usman,N., Beigelman,L. and Scott,W.G. (1998) The structural basis of hammerhead ribozyme self-cleavage. *Cell* **92**, 665–673.
20. Batey,R.T., Sagar,M.B. and Doudna,J.A. (2001) Structural and energetic analysis of RNA recognition by a universally conserved protein from the signal recognition particle. *J. Mol. Biol.*, **307**, 229–246.